Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

Maura R. Grossman*
Wachtell, Lipton, Rosen & Katz
mrgrossman@wlrk.com

ABSTRACT

Continuous active learning achieves high recall for technology-assisted review, not only for an overall information need, but also for various facets of that information need, whether explicit or implicit. Through simulations using Cormack and Grossman's TAR Evaluation Toolkit (SIGIR 2014), we show that continuous active learning, applied to a multi-faceted topic, efficiently achieves high recall for each facet of the topic. Our results assuage the concern that continuous active learning may achieve high overall recall at the expense of excluding identifiable categories of relevant information.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Search process, relevance feedback.

Keywords: Technology-assisted review; TAR; predictive coding; electronic discovery; e-discovery; test collections; relevance feedback; continuous active learning; CAL.

1. INTRODUCTION

The objective of technology-assisted review ("TAR"), first described in the context of electronic discovery ("eDiscovery") in legal matters [6], is to bring to the attention of a document reviewer substantially all relevant documents, and relatively few non-relevant ones, thereby maximizing recall and minimizing reviewer effort. The best reported results for TAR employ continuous active learning ("CAL"), in which a learning method presents the most-likely relevant documents to the reviewer in batches, the reviewer labels each document in each successive batch as relevant or not, and the labels are fed back to the learning method [6]. While CAL has been shown to achieve high recall with less effort than competing methods (including exhaustive manual review [7] and non-interactive supervised learning [6]), it has been suggested that CAL's emphasis on the most-likely relevant documents may bias it to prefer documents like the ones it finds first, causing it to fail to discover one or more important, but dissimilar, classes of relevant documents [11, 8].

In legal matters, an eDiscovery request typically comprises between several and several dozen requests for production ("RFPs"), each specifying a category of information sought. A review effort that fails to find documents relevant to each of the RFPs (assuming such documents exist) would likely be deemed deficient. In other domains, such as news services, topics are grouped into hierarchies, either explicit or implicit. A news-retrieval effort for "sports" that omits articles about "cricket" or "soccer" would likely be deemed inadequate, even if the vast majority of articles – about baseball, football, basketball, and hockey – were found. Similarly, a review effort that overlooked relevant short documents, spreadsheets, or presentations would likely also be seen as unsatisfactory.

We define a "facet" to be any identifiable subpopulation of the relevant documents, whether that subpopulation is defined by relevance to a particular RFP or subtopic, by file type, or by any other characteristic. Our objective is to determine whether CAL is able to achieve high recall over all facets, regardless of how they are identified. To this end, we used Cormack and Grossman's TAR Evaluation Toolkit ("Toolkit"),[1] grouping together the four RFPs supplied with the Toolkit as one overall topic, and treating each of the four RFPs as facets. We then computed recall as a function of review effort for the overall topic, as well as for each of the facets. We also computed recall separately for facets consisting of short documents, word-processing files, spreadsheets, and presentations.

We repeated our experiments, importing the Reuters RCV1-v2 dataset [10] into an adapted version of the Toolkit, using each of the RCV1-v2 top-level subject categories as an overall information need, and treating each of the 82 bottom-level subject categories as a facet.

2. TREC LEGAL TRACK EXPERIMENTS

In the Toolkit, we created a new overall topic, "all," by combining the topics supplied with the Toolkit (topics 201, 202, 203, and 207 from the TREC 2009 Legal Track [9]) as follows: As a seed set, we used 1,000 documents selected at random from the "seed query" hits in the Toolkit for the four topics [6, Table 3, p. 155]; as training and gold standards, we used the union of the respective standards supplied for the four topics, with the effect that, for both training and

*The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767771 .

[1] `http://cormack.uwaterloo.ca/cormack/tar-toolkit/`.

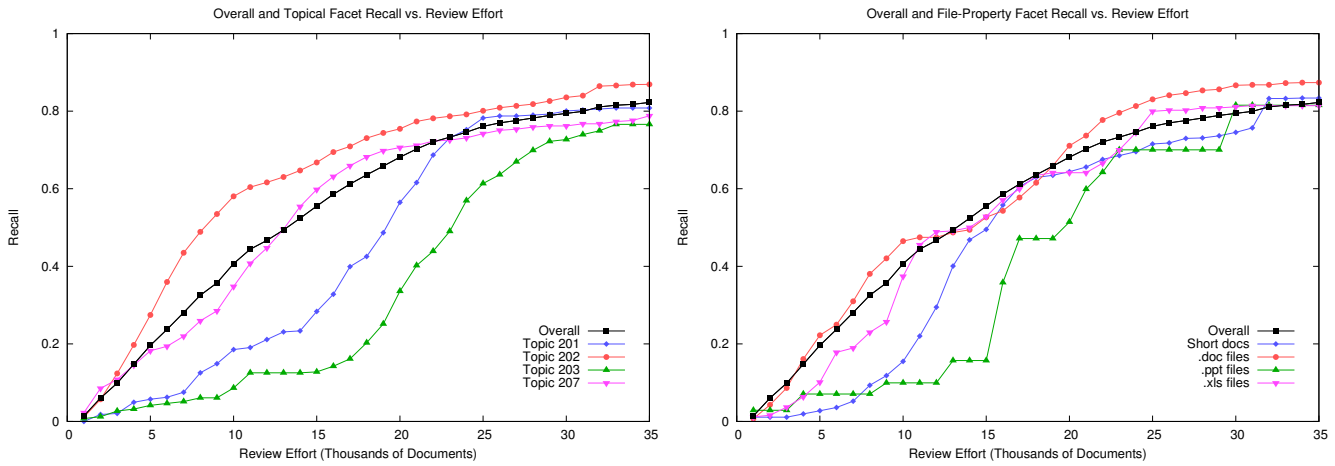

Figure 1: Overall and facet recall as a function of effort for the four TREC 2009 Legal Track topics. The left panel shows recall for an overall review effort, as well as recall with respect to each of four facets represented by the TREC topics. The right panel shows overall recall, as well as recall for facets represented by short documents, word-processing files, presentations, and spreadsheets.

evaluation, a document was considered relevant to the overall topic, "all," if it was relevant to any of the four facets represented by the Toolkit-supplied topics. We ran the "actkeysvm" CAL implementation, without modification, and captured the ordered list of documents presented for review. We computed recall for the "all" topic, and for each facet, at each position in the list.

We further evaluated recall with respect to four other facets: "short," ".doc," ".xls," and ".ppt," representing short documents ($< 1K$ bytes), word-processing files, spreadsheets, and presentations, respectively. Our results are shown in Figure 1. While it is apparent that, at the outset, topics 201 and 203, as well as short documents and presentations, lag behind; at high recall levels, there is little variance among the recall levels of the facets.

3. RCV1-V2 EXPERIMENTS

We next adapted the Toolkit to work with the RCV1-v2 dataset [10]. We used tf-idf Porter-stemmed word features, and SVM^{light}, following accepted practice [10]. We used as information needs the four top-level categories in the RCV1-v2 subject hierarchy, titled "corporate, industrial," "economics," "government and social," and "markets," with the corresponding subject codes, "CCAT," "ECAT," "GCAT," and "MCAT," respectively. As facets, we used the bottom-level categories in the RCV1-v2 subject hierarchy. We ignored the intermediate levels, as they are simple unions of the bottom-level categories. For seed queries, we used the titles of the top-level categories. We used the RCV1-v2 labels as both the training and gold standards.

Figure 2 shows overall and facet recall, as a function of review effort, for each of the four RCV1-v2 overall information needs. The results mirror those for the TREC 2009 dataset; however, a much higher level of recall is achieved, and convergence occurs as that higher level is approached. It appears that convergence coincides with a decline in marginal precision which, for these experiments, takes place in the neighborhood of 90% recall.

Among the results in Figure 2, one facet – "GMIL" – is an obvious outlier. Its recall reaches 80% only with double the review effort required for all other facets. In the RCV1-v2 collection, only five documents are labeled relevant to GMIL (bottom-level topic title: "Millennium Issues"). We examined these documents and found that four of them contain the phrase "millennium bug"; see, for example, the left panel of Table 1. We then searched the dataset and found 141 documents containing this same phrase, of which four were labeled relevant to GMIL, 48 were labeled relevant to some other facet of GCAT, and 93 were not labeled relevant to any facet. The right panel of Table 1 shows one such document. On reviewing these and other examples, we were unable to glean the criteria used to distinguish relevant from non-relevant documents, and thus attribute this outlier result to apparent mislabeling in the RCV1-v2 dataset.

4. WHEN TO STOP REVIEW

The objective of finding substantially all relevant documents suggests that CAL – or any other review effort – should continue until high recall has been achieved, and achieving higher recall would require disproportionate effort. Measuring recall is problematic, due to imprecision in the definition and assessment of relevance [3, 12, 8], and the effort, bias, and imprecision associated with sampling [2, 1, 8]. Accordingly, it is difficult to specify an absolute threshold value that constitutes "high recall," or to determine reliably that that such a threshold had been reached. Arguably, 75% to 80% recall would be sufficient for the TREC review detailed in Figure 1, because at that level, the recall for the facets is uniformly high, and a higher level can be achieved only with disproportionate effort. On the other hand, 90% or higher recall would be necessary to establish the adequacy of the RCV1-v2 reviews detailed in Figure 2. Only above 90% overall recall is the recall for facets uniformly high, and 90% recall is achievable with proportionate effort.

We suggest that, as an alternative to a fixed recall target, marginal precision might be a better indication of the completeness of a CAL review. In all of our experiments,

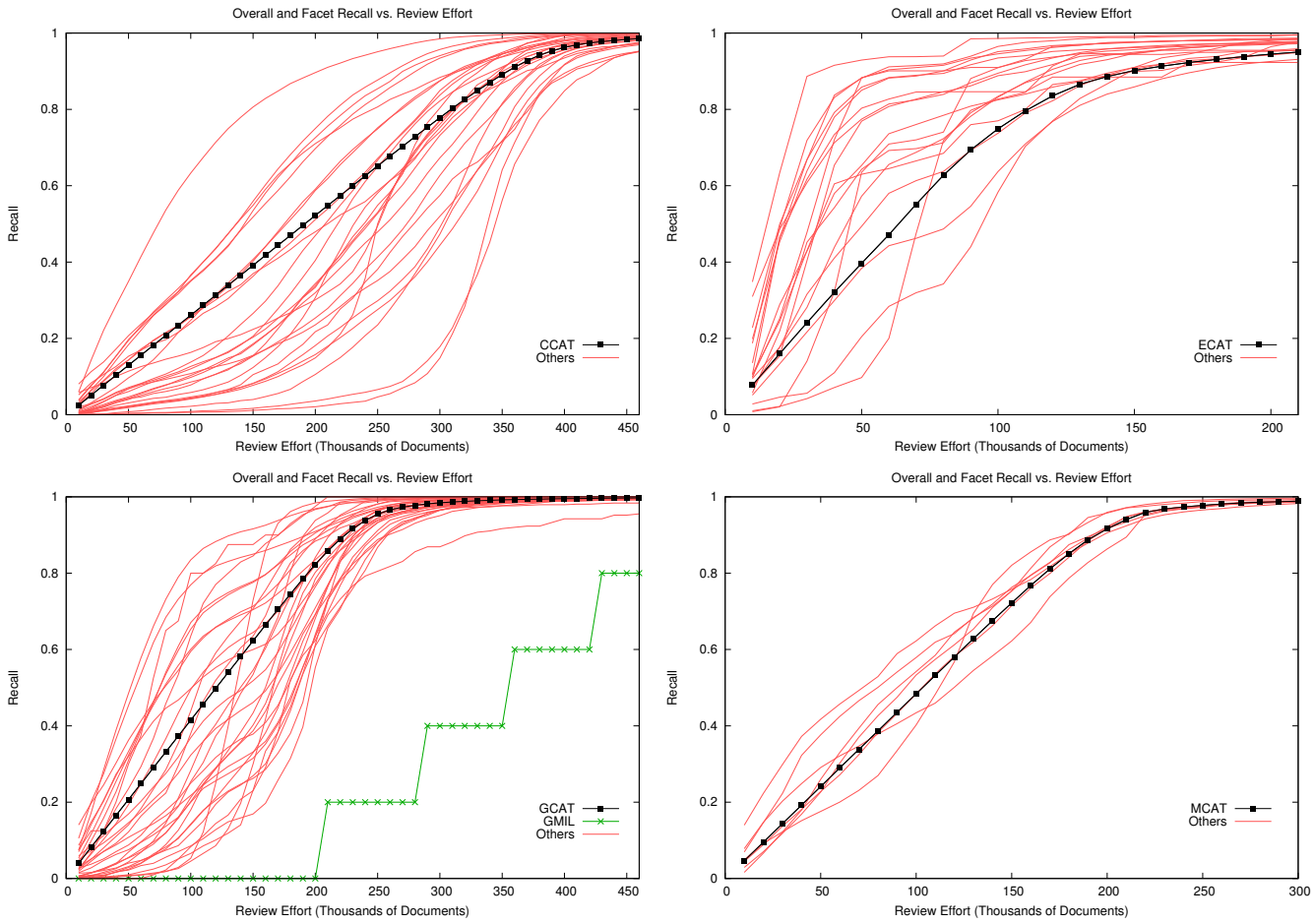

Figure 2: Overall and facet recall as a function of effort for the RCV1-v2 dataset.

we observed that the precision of each successive batch of 1,000 documents rose rapidly to nearly 100%, was sustained at nearly 100%, and then fell off. Table 2 illustrates that, in these experiments, stopping the review when marginal precision falls below one-tenth of its previously sustained value is a good predictor of high recall for the overall information need, as well as the facets, with proportionate effort.

5. DISCUSSION

A sign test shows our result to be significant ($p < 0.03$), by virtue of being observed for six of six separate experiments.

CAL is a greedy method that always chooses the most-likely relevant documents for review. It is to be expected that, at the outset, it chooses documents representing the easiest-to-identify facets, due to their subject matter, file properties, or abundance. As those documents are exhausted, others representing new facets become the most-likely relevant documents, until no more likely relevant documents remain. Only when the most-likely relevant documents from all facets have been exhausted, does marginal precision drop to a *de minimus* level.

While our findings suggest that it may be unnecessary, neither our theory of CAL's operation nor our results suggest that it would be harmful to train the learning method using additional seed documents – found by ad hoc means – to represent important facets that are are known to the reviewer at the outset, or become known during the course of the review process [5].

Our experiments suggest that when a review achieves sustained high precision, and then drops off substantially, one may have confidence that substantially all facets of relevance have been explored. In addition to offering a potentially better prediction of completeness, precision can be readily calculated throughout the review, while recall cannot. Further research is necessary to determine the extent to which marginal precision may afford a reliable quantitative estimate of review completeness, including coverage of different facets of relevance.

While sharing general motivation with efforts to achieve novelty and diversity in ad hoc retrieval [4], CAL seeks to achieve high recall rather than to reduce redundancy, and does so using a depth-first rather than breadth-first approach. We conducted an auxiliary experiment to investigate whether a strategy of using separate reviews for each facet would improve on the combined review strategy reported here. We found that the overall effort required to achieve 75% recall for every facet was higher for the separate review strategy. It remains to be seen whether other diversity-focused methods might improve on the purely depth-first results presented here.

Labeled "Relevant" in RCV1-v2	Labeled "Not relevant" in RCV1-v2
Okura up on millennium bug software demand. TOKYO 1997-08-15 Shares of Okura & Co Ltd surged on Friday afternoon due to the expectation that its software business would benefit from the so-called millennium bug problem. The stock was the top percentage gainer on the Tokyo Stock Exchange's first section in the afternoon session. Okura's shares were up 55 yen at 455 yen as of 0435 GMT. (c) Reuters Limited 1997-06-16	**Complete Business Solutions gets contract.** FARMINGTON HILLS, Mich. 1997-06-16 Complete Business Solutions Inc said early Monday that South Carolina Electric & Gas Co has awarded it a contract to manage and implement its Year 2000 code conversion project and deal with issues related to the "millennium bug". The project is expected to require changes to over three million lines of code at the utility and to be in the multi-million dollar range, Complete Business said. The company received the contract as part of a bidding process that included six other vendors. SCE&G is the principal subsidiary of SCANA Corp. ((– New York Newsdesk 212-859-1610)) (c) Reuters Limited 1997

Table 1: Conflicting relevance labels for category GMIL (bottom-level topic title: "Millennium Issues"). Of 141 documents in the RCV1-v2 collection containing the phrase "millennium bug," four were labeled "relevant," while 137 were labeled "not relevant."

	TREC	CCAT	ECAT	GCAT	GCAT (excl. GMIL)	MCAT
Review Effort (K-docs)	34	436	166	281	281	237
Overall Precision	0.587	0.856	0.664	0.848	0.848	0.840
Overall Recall	0.818	0.979	0.919	0.996	0.996	0.972
Overall F_1	0.684	0.913	0.771	0.916	0.916	0.901
Lowest Facet Recall	0.766	0.924	0.890	0.200	0.863	0.958

Table 2: Review effort and effectiveness when marginal precision (measured on the last batch of 1,000 documents) falls below 10%. Effort is measured in terms of thousands of documents reviewed; effectiveness is measured in terms of overall precision, recall, and F_1, as well as the lowest recall obtained for any facet.

6. CONCLUSION

For all experiments, our results are the same: CAL achieves high overall recall, while at the same time achieving high recall for the various facets of relevance, whether topics or file properties. While early recall is achieved for some facets at the expense of others, by the time high overall recall is achieved – as evidenced by a substantial drop in overall marginal precision – all facets (except for a single outlier case that we attribute to mislabeling) also exhibit high recall. Our findings provide reassurance that CAL can achieve high recall without excluding identifiable categories of relevant information.

7. REFERENCES

[1] M. Bagdouri, D. D. Lewis, and D. W. Oard. Sequential testing in classifier evaluation yields biased estimates of effectiveness. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 933–936, 2013.

[2] M. Bagdouri, W. Webber, D. D. Lewis, and D. W. Oard. Towards minimizing the annotation cost of certified text classification. In *Proceedings of the 22nd ACM International Conference Information and Knowledge Management*, pages 989–998, 2013.

[3] D. C. Blair. STAIRS redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47(1):4–22, Jan. 1996.

[4] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkann, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–666, 2008.

[5] G. Cormack and M. Mojdeh. Machine learning for information retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks. In *Eighteenth Text REtrieval Conference*, 2009.

[6] G. V. Cormack and M. R. Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–162, 2014.

[7] M. R. Grossman and G. V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17(3):1–48, 2011.

[8] M. R. Grossman and G. V. Cormack. Comments on "The implications of rule 26(g) on the use of technology-assisted review". *Federal Courts Law Review*, 7:285–313, 2014.

[9] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the TREC 2009 Legal Track. In *The Eighteenth Text REtrieval Conference*, 2009.

[10] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[11] K. Schieneman and T. Gricks. The implications of Rule 26(g) on the use of technology-assisted review. *Federal Courts Law Review*, 7(1):239–274, 2013.

[12] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.